Special Issue Paper

Environmetrics

Received: 4 April 2014, Revised: 15 August 2014, Accepted: 17 August 2014, Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/env.2305

On the impact of covariate measurement error on spatial regression modelling

Md Hamidul Huque[a], Howard D. Bondell[b] and Louise Ryan[a]*

Spatial regression models have grown in popularity in response to rapid advances in geographic information system technology that allows epidemiologists to incorporate geographically indexed data into their studies. However, it turns out that there are some subtle pitfalls in the use of these models. We show that the presence of covariate measurement error can lead to significant sensitivity of parameter estimation to the choice of spatial correlation structure. We quantify the effect of measurement error on parameter estimates and then suggest two different ways to produce consistent estimates. We evaluate the methods through a simulation study. These methods are then applied to data on ischaemic heart disease. Copyright © 2014 John Wiley & Sons, Ltd.

Keywords: attenuation; environmental epidemiology; geostatistics; measurement error; mixed models; random effects; SEIFA; sensitivity; spatial correlation; spatial linear regression

1. INTRODUCTION

Advances in statistical methodology, together with the availability of geographically referenced health databases, present a unique opportunity to investigate the environmental, social and behavioural factors underlying geographic variations (Elliott and Wartenberg, 2004). In health research, for example, social epidemiologists seek to assess the impact of socio-demographic characteristics of a community on the health of individuals living in that community (Elliot *et al.*, 2000). The analysis of geo-coded data is complicated by correlations among observations located near each other. Regression analysis ignoring these spatial correlations leads to incorrect inference on the estimated regression coefficients by narrowing confidence intervals. Mixed-effect models provide a convenient way of modelling spatial correlations by incorporating random effects with a spatial correlation structure (Waller and Gotway, 2004). In this paper, we focus on how such models perform when covariates of interest are measured with error.

In the case study that motivates this paper, Australian researchers explored the relationship between the Socio-Economic Indexes for Areas (SEIFA) (an area-based measure of socio-economic status produced by the Australian Bureau of Statistics) and acute hospitalization for ischaemic heart disease (IHD) for approximately 600 postcodes in New South Wales, Australia (Burden *et al.*, 2005; Guha *et al.*, 2009). Regression models suggest a strong association between SEIFA and IHD, even after adjusting for factors such as age, gender, population density and other factors that might influence the outcome. However, exploratory analysis reveals that the estimated coefficient of the SEIFA index from such models depends strongly on the assumed spatial correlation structure. Briefly, the estimated SEIFA coefficients are all significantly negative, confirming that IHD rates decrease as social advantage increases. However, the magnitude of the effect varies by more than a factor of 2, depending on whether a spatial correlation adjustment is made. Similar sensitivity to assumed spatial correlation structure can be seen in the analysis of the well-known Scottish Lip Cancer data (Breslow and Clayton, 1993; Clayton *et al.*, 1993). In another spatial epidemiological study, Molitor *et al.* (2007) fit a model for the effect of NO_2 exposure on lung function. They considered a series of models including one based on a conditional autoregressive model. They observed that models with a spatial structure give smaller effect estimates as compared with models without a spatial structure. These results suggest that estimated coefficients from a spatial regression model can be highly sensitive to whether and how spatial variation is accommodated. In this paper, we show that such sensitivity is especially likely to occur when the covariate of interest has been measured with error.

The presence of measurement error in the covariate of interest arises in many epidemiological and socio-behavioural studies. For example, in the study of geographical variation in bladder cancer rates, lung cancer risk might be included in the model as a proxy for smoking exposure (Clayton *et al.*, 1993). In environmental epidemiology, individual air pollution exposures might be approximated by the distance from the polluted sites or by using the measures at a few monitoring sites (Carroll *et al.*, 1997). Further examples include geographical studies relating cancer incidence and mortality to dietary intakes (Cook and Pocock, 1983; Prentice and Sheppard, 1990).

* *Correspondence to: Louise Ryan, School of Mathematical Sciences, University of Technology, Sydney, NSW 2007, Australia. E-mail: Louise.M.Ryan@uts.edu.au*

a *School of Mathematical Sciences, University of Technology, Sydney, NSW 2007, Australia*

b *Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203, U.S.A.*

Many papers have appeared in the literature over the years on covariate measurement error in the context of independent data (Carroll *et al.*, 2006; Fuller, 1987; Wansbeek and Meijer, 2000; Ruppert *et al.*, 2009). In case of linear regression with independent data, it is well known that the presence of exposure measurement error causes estimated regression coefficients to attenuate towards the null. However, relatively few have addressed the effect of exposure measurement error in the context of correlated data with spatial structure. In epidemiological studies of association between air pollutants and health outcome, typically, data are available from few monitoring sites. Therefore, the measured exposure used in the analysis might be different from the underlying true exposure.

Xia and Carlin (1998) presented a spatio-temporal analysis of spatially correlated data with errors in the covariates, in the context of disease mapping. The authors empirically studied several alternative measurement error models using a Metropolis–Gibbs algorithm. Li *et al.* (2009) derived asymptotic bias expressions for estimated regression coefficients in the context of a spatial linear mixed model. They showed that the regression estimates obtained from the naïve use of error-prone covariates attenuate the estimated regression coefficient and that variance component estimates are inflated. They proposed the use of a maximum likelihood approach based on the EM algorithm to adjust for measurement error under the assumed error structure. However, their simulation assumes that the measurement error variance is known, and they did not assess the performance of their method in the case of misspecification. Their approach is also subject to a high computational burden and may lead to spurious results in the presence of outliers or model misspecification (Gryparis *et al.*, 2009; Szpiro *et al.*, 2011). Furthermore, Szpiro *et al.* (2011) argued that in the presence of spatial correlation, joint modelling becomes challenging as it is very difficult to separate the spatial correlation between exposure and outcome.

In this paper, we explore the sensitivity of estimated regression coefficients in spatial regression models, showing that it arises in settings where the covariate of interest has been measured with error. We show that ignoring measurement error attenuates estimated regression coefficients and observe that estimates can be very sensitive to the choice of assumed correlation structure in the model formulation. We derive expressions for the bias when measurement error is ignored and present some technical derivations that characterize the bias as a function of the degree of measurement error as well as the degree of spatial correlation in the covariate of interest and in the residuals. We show that the bias due to attenuation depends on the spatial correlation structure. When there is no or the same degree of spatial correlation in both the covariate and the measurement error, the bias in spatial linear model reduces to the familiar attenuation factors under OLS modelling of independent data, namely $\rho = \sigma_X^2/(\sigma_X^2 + \sigma_U^2)$, where σ_X^2 is the variance of the true covariate and σ_U^2 is the variance of the measurement error.

Based on these expressions, we propose two different strategies for obtaining consistent estimates: (i) adjusting the estimates using an estimated attenuation factor; and (ii) using an appropriate transformation of the error-prone covariate. We then evaluate the performance of these two approaches via simulations. These approaches do not require complex programming and can be implemented via readily available mixed-model software. Furthermore, we suggest ways to estimate measurement error variance from the data, rather than assuming measurement error variance as a known quantity. Our simulation results show that bias correction methods using the estimate of the measurement error work reasonably well in obtaining consistent estimates. However, estimation of the measurement error variance requires additional data or assumptions related to the underlying measurement error process. In the case of spatial epidemiology, validation data are typically rare. Therefore, we suggest employing a sensitivity analysis when dealing with measurement error problems in practice. We illustrate the methods using data on IHD and conclude with some practical guidelines.

2. MODEL FORMULATION

Suppose that X_i represents the true covariate of interest for spatial location i, $i = 1, \ldots, n$, and suppose that it is related to an outcome Y_i according to a linear model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{1}$$

where $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T \sim N(0, \boldsymbol{\Sigma}_\epsilon)$ and $\boldsymbol{\Sigma}_\epsilon$ is a covariance matrix, for now kept arbitrary. Let W_i be the observed covariate for spatial location i, related to the true covariate according to a classical measurement error model:

$$W_i = X_i + U_i$$

where $\boldsymbol{U} = (U_1, \ldots, U_n)^T \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_U)$. When $\boldsymbol{X} = (X_1, \ldots, X_n)^T$ is also normally distributed (say with mean $\boldsymbol{\mu}_X$ and covariance $\boldsymbol{\Sigma}_X$), straightforward algebra establishes that $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ and $\boldsymbol{W} = (W_1, \ldots, W_n)^T$ have a multivariate normal distribution,

$$\begin{pmatrix} \boldsymbol{Y} \\ \boldsymbol{W} \end{pmatrix} \sim MVN\left(\begin{pmatrix} (\beta_0 + \beta_1 \boldsymbol{\mu}_X)\boldsymbol{1} \\ \boldsymbol{\mu}_X \boldsymbol{1} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_\epsilon + \beta_1^2 \boldsymbol{\Sigma}_X & \beta_1 \boldsymbol{\Sigma}_X \\ \beta_1 \boldsymbol{\Sigma}_X & \boldsymbol{\Sigma}_X + \boldsymbol{\Sigma}_U \end{pmatrix} \right)$$

where $\boldsymbol{1}$ is an $n \times 1$ vector of 1's. Standard theory for the multivariate normal establishes that $\boldsymbol{Y}|\boldsymbol{W}$ is normally distributed with conditional mean

$$E(\boldsymbol{Y}|\boldsymbol{W}) = \beta_0 \boldsymbol{1} + \beta_1 (\boldsymbol{I} - \boldsymbol{\Lambda}) \boldsymbol{\mu}_x + \beta_1 \boldsymbol{\Lambda} \boldsymbol{W} \tag{2}$$

and conditional variance

$$Var(\boldsymbol{Y}|\boldsymbol{W}) = \boldsymbol{\Sigma}_\epsilon + \beta_1^2(\boldsymbol{I} - \boldsymbol{\Lambda})\boldsymbol{\Sigma}_X$$

where

$$\boldsymbol{\Lambda} = \boldsymbol{\Sigma}_X(\boldsymbol{\Sigma}_X + \boldsymbol{\Sigma}_U)^{-1} \tag{3}$$

For ease of discussion, assume that the variable $\boldsymbol{X}$ has been centred so that $\boldsymbol{\mu}_X = \boldsymbol{0}$. In direct analogy with standard measurement error settings, these results suggest that regression coefficients obtained by regressing the outcome ($\boldsymbol{Y}$) on the observed, but error-prone, covariate ($\boldsymbol{W}$) will lead to bias as well as inaccurate variance modelling. We proceed now to explore the nature of this bias under varying assumptions about the correlation structure for $\boldsymbol{Y}$ and $\boldsymbol{X}$ and the measurement error term, $\boldsymbol{U}$.

3. ASYMPTOTIC BIAS ANALYSIS

Suppose we fit model (1), naively replacing $\boldsymbol{X}$ with the error-prone version of the covariate $\boldsymbol{W}$ and assuming independence of the error terms in the model on $\boldsymbol{Y}$. The ordinary least squares estimate of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}^{ols} = \left(\boldsymbol{W}_*^T\boldsymbol{W}_*\right)^{-1}\boldsymbol{W}_*^T\boldsymbol{Y} \tag{4}$$

where $\boldsymbol{W}_*$ is the $n \times 2$ matrix with elements of the first column all equal to 1 and those of the second column corresponding to the $n \times 1$ vector $\boldsymbol{W}$. Under the true model and assuming $\boldsymbol{\mu}_X = 0$, it is straightforward to show that the limiting value of this estimate is

$$\begin{aligned}\tilde{\boldsymbol{\beta}}^{ols} &= \begin{pmatrix} 1 & 0 \\ 0 & trace\,(\boldsymbol{\Sigma}_X + \boldsymbol{\Sigma}_U) \end{pmatrix}^{-1} \begin{pmatrix} 1 & 0 \\ 0 & trace\,(\boldsymbol{\Sigma}_X) \end{pmatrix} \boldsymbol{\beta} \\ &= \begin{pmatrix} 1 & 0 \\ 0 & \rho^{ols} \end{pmatrix} \boldsymbol{\beta}\end{aligned}$$

where $\rho^{ols} = trace\,(\boldsymbol{\Sigma}_X)\,/trace\,(\boldsymbol{\Sigma}_X + \boldsymbol{\Sigma}_U)$; see the Appendix.

Using basic properties of the trace function, this simple formula leads to a number of interesting observations. For example, suppose that both $\boldsymbol{\Sigma}_X$ and $\boldsymbol{\Sigma}_U$ have constant diagonal elements σ_X^2 and σ_U^2, respectively, then the bias factor can be written as $\rho^{ols} = \sigma_X^2/(\sigma_X^2 + \sigma_U^2)$. This is the standard measurement error result (Carroll *et al.* 2006), namely that the estimated regression coefficient is biassed towards the null by an attenuation factor that reflects the proportion of the variability in the observed covariate $\boldsymbol{W}$, explained by the true covariate $\boldsymbol{X}$. Note that there is no bias in the estimated intercept in this case because we have assumed that $\boldsymbol{X}$ has zero mean. It is interesting to note that the result holds regardless of the correlation structures on the error term, $\boldsymbol{\Sigma}_\epsilon$.

In the next section, we consider the bias associated with fitting a generalized least squares (GLS) model in the presence of covariate measurement error. We will see that in this case, the degree of bias also depends on the assumed error structure.

3.1. Generalized least squares (GLS)

Suppose we obtain a GLS estimator of $\boldsymbol{\beta}$, under that assumption that the error term $\boldsymbol{\epsilon}$ has covariance matrix $\boldsymbol{\Sigma}_a$, with the subscript a denoting 'assumed'. For fixed $\boldsymbol{\Sigma}_a$, the estimator is

$$\hat{\boldsymbol{\beta}}^{gls} = \left(\boldsymbol{W}_*^T\boldsymbol{\Sigma}_a^{-1}\boldsymbol{W}_*\right)^{-1}\boldsymbol{W}_*^T\boldsymbol{\Sigma}_a^{-1}\boldsymbol{Y} \tag{5}$$

In the limit, under the true model and following similar arguments as in the OLS case, this estimate converges in probability to

$$\begin{aligned}\tilde{\boldsymbol{\beta}}^{gls} &= \begin{pmatrix} n & 0 \\ 0 & trace\left[\boldsymbol{\Sigma}_a^{-1}\,(\boldsymbol{\Sigma}_X + \boldsymbol{\Sigma}_U)\right] \end{pmatrix}^{-1} \begin{pmatrix} n & 0 \\ 0 & trace\left[\boldsymbol{\Sigma}_a^{-1}\boldsymbol{\Sigma}_X\right] \end{pmatrix} \boldsymbol{\beta} \\ &= \begin{pmatrix} 1 & 0 \\ 0 & \rho^{gls} \end{pmatrix} \boldsymbol{\beta}\end{aligned}$$

where $\rho^{gls} = trace[\boldsymbol{\Sigma}_a^{-1}\boldsymbol{\Sigma}_X]/trace[\boldsymbol{\Sigma}_a^{-1}(\boldsymbol{\Sigma}_X + \boldsymbol{\Sigma}_U)]$.

As in the OLS case, this simple formula also yields a number of interesting observations with important practical implications. First of all, because we can write

$$\rho^{gls} = \left[1 + trace\left(\boldsymbol{\Sigma}_a^{-1}\boldsymbol{\Sigma}_U\right)/trace\left(\boldsymbol{\Sigma}_a^{-1}\boldsymbol{\Sigma}_X\right)\right]^{-1} \tag{6}$$

it follows that there will always be an attenuation of the estimated regression coefficient towards the null.

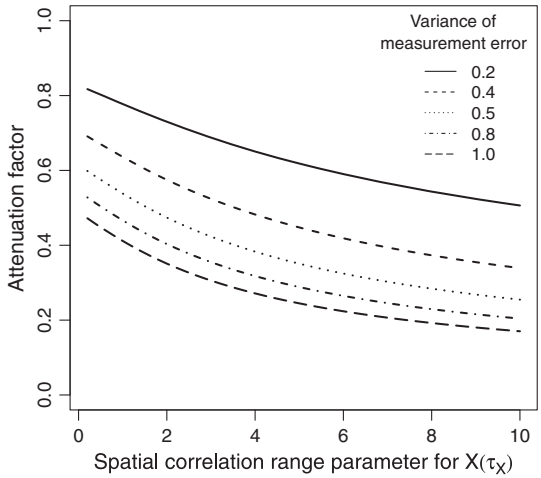

Figure 1. Attenuation factor associated with varying degree of measurement error

Figure 1 shows the attenuation factor associated with fitting GLS, ρ^{gls}, under the assumption that $\boldsymbol{X}$ and $\boldsymbol{\epsilon}$ have unit variance and an exponential spatial covariance structure with the correlation between two observations; a distance h units apart is given by $Cor(h) = \exp(-h/\tau)$, where τ denotes the range.

Each line in Figure 1 corresponds to a unique value of σ_U^2, the measurement error variance. The x-axis in the figure varies according to the value of the range parameter τ_x, which reflects the strength of the spatial correlation in the true covariate, $\boldsymbol{X}$. All calculations in the figure assume that there is zero spatial correlation in the measurement error term, $\boldsymbol{U}$. Note that, as the range parameter goes to zero ($\tau_x \to 0$), the attenuation factor becomes identical to that which would be obtained if OLS were used instead of GLS. Of course, these results could change in the presence of other covariates in the model (Zeger *et al.* 2000; Schwartz and Coull 2003).

From (6), it is clear that the two attenuation factors, ρ^{gls} and ρ^{ols}, will equal the familiar attenuation factor under OLS modelling for independent data, namely $\rho = \sigma_X^2/(\sigma_X^2 + \sigma_U^2)$, under a variety of circumstances, including the following:

1. $\boldsymbol{\Sigma}_X = \sigma_X^2 \boldsymbol{I}$ and $\boldsymbol{\Sigma}_U = \sigma_U^2 \boldsymbol{I}$. That is, there is no spatial correlation in X or U, and both random variables have homogenous variance.
2. When the degree of spatial correlation is the same for $\boldsymbol{X}$ and the measurement error, $\boldsymbol{U}$. That is, $\boldsymbol{\Sigma}_X = \sigma_X^2 \boldsymbol{R}$ and $\boldsymbol{\Sigma}_U = \sigma_U^2 \boldsymbol{R}$, where $\boldsymbol{R}$ is a spatial correlation matrix.

Note that these results hold regardless of the value of $\boldsymbol{\Sigma}_a$, the assumed correlation for the residuals in regression models. In practice, ρ^{gls} and ρ^{ols} may differ depending on how well the assumed spatial correlation structure resembles the true process of the underlying covariance structure. In the next section, we propose several approaches to adjust for measurement error in spatial regression settings.

4. BIAS CORRECTION

In the previous section, we have shown that the presence of measurement error in covariates attenuates estimated regression coefficient to the null. A consistent estimate of the true regression coefficient can be obtained if we can estimate the various parameters that govern the measurement error process. This is possible if we have access to a validation data set without measurement error (Carroll *et al.*, 2006). In the context of spatial epidemiology, however, validation data are rarely available. Therefore, we need additional assumptions to estimate the components of the attenuation factor. Without such assumptions or validation data, the measurement error and the true residual error variance are not identifiable in both cases. In this paper, we considered two different sets of assumptions that lead to the model identifiability. The first approach assumes that the true covariate $\boldsymbol{X}$ is smooth and that any observed nugget effect must be a measurement error. The second assumes that measurement error variance is fixed and known over a feasible range. A sensitivity analysis is then carried out over the feasible range of known measurement error variance. Similar to Li *et al.* (2009), we assume that the underlying covariate process $\{X_i\}$ defined in Section 2 contains all the spatial correlation and that the measurement error is pure noise, that is, $\boldsymbol{\Sigma}_U = \sigma_U^2 \boldsymbol{I}$. Under this assumption, the attenuation factor from (6) becomes

$$\rho^{gls} = \left[1 + \sigma_U^2 trace\left(\boldsymbol{\Sigma}_a^{-1}\right)/trace\left(\boldsymbol{\Sigma}_a^{-1}\boldsymbol{\Sigma}_X\right)\right]^{-1} \tag{7}$$

The OLS version can be obtained from the special case where $\boldsymbol{\Sigma}_a = \sigma_\epsilon^2 \boldsymbol{I}$.

We examine two different bias correction strategies to obtain a consistent estimate of the regression coefficient. The first approach deals with estimation of each of the components of ρ^{gls}, and the second uses a linear transformation of the error-prone covariate, $\boldsymbol{W}$.

Both methods require knowledge of $\mathbf{\Sigma}_X$ and σ_U^2 or their estimated values. We estimated $\mathbf{\Sigma}_X$ and σ_U^2 by fitting the error-prone covariate ($\boldsymbol{W}$) in an intercept-only model with an assumed spatial correlation structure. Under the assumption that measurement error is pure noise and $\mathbf{\Sigma}_X$ is a smooth spatial covariate with no nugget, the aforementioned model gives us a maximum likelihood estimate of the nugget effect in $\boldsymbol{W}$, which corresponds to σ_U^2. Similarly, fitting $\boldsymbol{Y}$ on $\boldsymbol{W}$ with a spatial correlation structure gives us a maximum likelihood estimate of the underlying residual covariance structure, $\mathbf{\Sigma}_\epsilon$. The first method additionally requires an estimate of $\mathbf{\Sigma}_a$.

4.1. Method I: method of moments

This method involves post-analysis adjustment of the estimated regression coefficient using an estimate of the attenuation factor. Ignoring the measurement error and performing a likelihood analysis under the assumed covariance structure of $\boldsymbol{Y}|\boldsymbol{X}$ using $\boldsymbol{W}$ instead of $\boldsymbol{X}$ result in estimates denoted by $\hat{\boldsymbol{\beta}}^{ols}$ or $\hat{\boldsymbol{\beta}}^{gls}$ depending on whether OLS or GLS has been used. Let $\hat{\beta}_1$ be the estimate of the corresponding slope from the regression (1), where for ease of exposition, we leave off the superscript *ols* or *gls*. We have shown that its limiting value is $\rho\beta_1$. Denote its variance by σ_*^2. We then define an adjusted estimate $\hat{\beta}_1^{adj} = \hat{\beta}_1/\hat{\rho}$, where $\hat{\rho}$ is an estimate of the attenuation factor defined in (7) and where the estimated variance $\hat{Var}(\hat{\beta}_1^{adj}) = \hat{\rho}^{-2}\sigma_*^2$. An estimate of ρ is obtained by substituting $\hat{\sigma}_U^2$, $\hat{\mathbf{\Sigma}}_X$ and $\hat{\mathbf{\Sigma}}_a$ in (7).

4.2. Method II: transformation method

Recall from (2) that $E(\boldsymbol{Y}|\boldsymbol{W}) = \beta_0\mathbf{1} + \beta_1(\boldsymbol{I} - \mathbf{\Lambda})\boldsymbol{\mu}_X + \beta_1\mathbf{\Lambda}\boldsymbol{W}$, where $\mathbf{\Lambda} = \mathbf{\Sigma}_X(\mathbf{\Sigma}_X + \mathbf{\Sigma}_U)^{-1}$. This suggests the use of a linear transformation of $\boldsymbol{W}$ to achieve an appropriate linear regression model that can be fitted to yield a consistent estimate of $\boldsymbol{\beta}$. Specifically, letting $\boldsymbol{T} = \boldsymbol{\mu}_X + \mathbf{\Lambda}(\boldsymbol{W} - \boldsymbol{\mu}_X)$, it follows that $\boldsymbol{T} \sim (\boldsymbol{\mu}_X, \mathbf{\Lambda}\mathbf{\Sigma}_X)$ and $Cov(\boldsymbol{Y}, \boldsymbol{T}) = \beta_1\mathbf{\Lambda}\mathbf{\Sigma}_X$. Hence, using the joint normality of $\boldsymbol{W}$ and $\boldsymbol{Y}$, we have $E(\boldsymbol{Y}|\boldsymbol{T}) = \beta_0 + \beta_1\boldsymbol{T}$ and $Var(\boldsymbol{Y}|\boldsymbol{T}) = Var(\boldsymbol{Y}|\boldsymbol{W}) = \Sigma_\epsilon + \beta_1^2(\boldsymbol{I} - \mathbf{\Lambda})\mathbf{\Sigma}_X$.

Define $\tilde{\boldsymbol{T}}$ as the estimator of $\boldsymbol{T}$, obtained by substituting in consistent estimates of $\boldsymbol{\mu}_X$ and $\mathbf{\Lambda}$. The outcome $\boldsymbol{Y}$ can then be regressed on $\tilde{\boldsymbol{T}}$, with an assumed spatial correlation structure, via a linear mixed model to obtain a consistent estimate of β_1 and corresponding standard error.

5. SIMULATIONS

We conducted a simulation study to evaluate the finite-sample properties of two methods proposed in the previous section to adjust for measurement error. We simulated 100 sample locations randomly within a $d \times d$ rectangular grid, where d is taken to have a value of either 40 or 80. Specifically, the ith random sample location s_i was generated by simulating two coordinates (e.g. latitude and longitude) from a Uniform$[0, d]$ distribution. Given the set of s_i's, the unobserved true covariate $\boldsymbol{X}$ was generated with mean 0 and covariance matrix $\mathbf{\Sigma}_X$, where $\mathbf{\Sigma}_X$ was assumed to have an exponential correlation structure with unit variance. This implies that the correlation between two observations distance h units apart is $(1 - \eta_x) * \exp(-h/\tau_x)$, where τ_x is the range parameter and η_x characterizes the so-called nugget effect. We considered three different range parameters ($\tau_x = 1, 5, 10$) resulting in minimal, moderate and high correlations among the values of X's with a nugget effect of $\eta_x = 0.1$.

The observed error-prone versions, $\boldsymbol{W}$, of the true covariate were generated by adding Gaussian noise with variance σ_U^2 to $\boldsymbol{X}$. Outcome data, $\boldsymbol{Y}$, were then generated according to (1), the slope and intercept parameter are taken as $(\beta_0, \beta_1)^T = (1, 2)^T$ and the error variances were generated using a similar exponential correlation structure as $\mathbf{\Sigma}_X$, but with different range parameters. We also add a random Gaussian noise to the residual error variance (nugget effect). The variance parameter and the nugget for the residual error were taken as 0.5 and 0.1, respectively. We used the nlme package (Pinheiro *et al.*, 2013) in generating simulated data with exponential spatial correlation and in model fitting. Then we extracted the covariate matrices from the object of lme fit using the mgcv package (Wood, 2006) in R (R Core Team, 2013).

To study the performance of our proposed methods under various degrees of correlation within the rectangular grid and for various values of the measurement error variance, we simulated data based on various combinations of measurement error variances ($\sigma_U^2 = 0.0$, 0.2 and 0.5). To simplify our presentation, only the results with measurement error variance $\sigma_U^2 = 0.2$ in 80×80 grid scales are illustrated. In general, the results obtained by varying the measurement error and/or size of the grid are similar.

Table 1 shows the average of the estimated regression coefficients, empirical standard errors and average of the estimated standard errors under nine different combinations of spatial correlation in the covariate $\boldsymbol{X}$ and in the error for the model $\boldsymbol{Y}$ given $\boldsymbol{X}$, based on 1000 simulations. The first column of Table 1 specifies the combination of range parameters (τ_X, τ_ϵ) used in that particular simulation. The second and third columns show the estimated regression parameters under OLS based on using the true covariate $\boldsymbol{X}$ and the error-prone covariates $\boldsymbol{W}$, respectively (equation (4)). The fourth column shows the results from fitting a linear mixed model (using the *lme* function in R) with assumed exponential correlation structure, but without adjusting for measurement error. The fifth and sixth columns present the bias-corrected estimates of the regression parameter $\boldsymbol{\beta}_1$ using Methods I and II, respectively. The next two columns of the table represent the results from Methods I and II when true measurement error variances were used instead of estimated values. That is, results in Column 5 use the estimated measurement error, and those in Column 7 use the true value of the measurement error variance under Method I. Similarly, Columns 6 and 8 represent the results obtained using Method II based on the *estimated* and *true* measurement error variances, respectively. The last column of the table shows results from Method II when all the components of $\mathbf{\Lambda}$ were calculated using true values (i.e. values used in data generation).

The simulation results confirmed that the degree of bias for linear mixed model with error-prone covariate vary with the strength of the spatial correlation structure of covariate as well as residuals. However, our proposed bias correction methods perform well in terms of providing consistent estimates of the regression coefficient. Both methods underestimate the true regression coefficient when measurement error is estimated and there is very low correlation in $\boldsymbol{X}$. This makes sense because the nugget effect in $\boldsymbol{X}$ is non-identifiable in that setting. This is because the assumption that the true covariate $\boldsymbol{X}$ is smooth is no longer valid; hence, estimates are not reliable in such situations.

Table 1. Simulation results using different combinations of range parameters

				Bias-corrected *lme*				
	OLS		*lme*	Using estimated σ_u^2		Using true σ_u^2		Using true $\mathbf{\Lambda}$
Range[a] (τ_X, τ_ϵ)	Using X	Using W	Using W	Method I	Method II	Method I	Method II	Method II
				Estimated coefficient				
(1, 1)	1.999	1.689	1.683	1.867	1.838	2.039	2.000	1.995
(1, 5)	2.000	1.692	1.682	1.886	1.844	2.048	2.001	1.997
(1, 10)	2.001	1.691	1.681	1.889	1.849	2.038	2.001	1.995
(5, 1)	1.999	1.682	1.665	2.075	1.987	2.039	1.990	1.995
(5, 5)	2.002	1.687	1.641	2.106	1.998	2.051	1.988	1.998
(5, 10)	2.004	1.687	1.630	2.106	2.000	2.039	1.986	1.997
(10, 1)	2.000	1.666	1.638	2.113	2.013	2.040	1.990	1.996
(10, 5)	2.002	1.668	1.584	2.151	2.028	2.050	1.984	1.996
(10, 10)	2.005	1.670	1.562	2.173	2.048	2.037	1.982	1.997
				Empirical standard error				
(1, 1)	0.075	0.097	0.099	0.473	0.321	0.138	0.126	0.115
(1, 5)	0.077	0.099	0.099	0.614	0.331	0.147	0.127	0.115
(1, 10)	0.077	0.099	0.095	0.676	0.349	0.142	0.123	0.112
(5, 1)	0.079	0.104	0.107	0.583	0.409	0.145	0.131	0.120
(5, 5)	0.091	0.110	0.113	0.753	0.508	0.178	0.139	0.123
(5, 10)	0.098	0.116	0.115	0.768	0.512	0.172	0.143	0.127
(10, 1)	0.083	0.114	0.121	0.494	0.334	0.176	0.139	0.125
(10, 5)	0.102	0.126	0.142	0.616	0.418	0.247	0.159	0.137
(10, 10)	0.117	0.134	0.145	0.692	0.469	0.210	0.165	0.142
				Average of estimated standard errors				
(1, 1)	0.075	0.100	0.099	0.110	0.109	0.120	0.119	0.118
(1, 5)	0.074	0.100	0.096	0.108	0.107	0.118	0.115	0.114
(1, 10)	0.073	0.099	0.093	0.105	0.104	0.113	0.112	0.110
(5, 1)	0.076	0.101	0.101	0.127	0.124	0.125	0.120	0.120
(5, 5)	0.075	0.100	0.101	0.130	0.126	0.126	0.121	0.121
(5, 10)	0.073	0.099	0.098	0.127	0.124	0.123	0.119	0.119
(10, 1)	0.078	0.103	0.104	0.135	0.129	0.131	0.124	0.123
(10, 5)	0.077	0.103	0.106	0.145	0.137	0.138	0.129	0.129
(10, 10)	0.075	0.102	0.104	0.146	0.139	0.137	0.128	0.128

Reported numbers are averaged over 1000 simulations with 100 observations per simulation with measurement error variance 0.2.

[a](τ_X, τ_ϵ)values of the range parameter following exponential correlation in $\boldsymbol{X}$ and the error term in the model on $\boldsymbol{Y}$, respectively.

To assess the sensitivity of the true spatial correlation structure on parameter estimation, we run a simulation with misspecified spatial correlation structure. In this simulation, we generated data using an exponential covariance structure, but fitted under the assumption of a Gaussian covariance structure. Figure 2 shows the distribution of estimated coefficients when estimated and true values of σ_u^2 are used with Methods I (a–b) and II (c–d), under different range parameters combined with true and misspecified covariate structures. For each combination of range parameters, the first boxplot (from left) represents results from the misspecified covariance structure. The results obtained for the other combination of range parameters (not shown in this figure) are similar.

Our simulation results illustrate that proposed methods are quite robust in case of misspecification of underlying covariance structure. However, the accuracy of the methods depends largely on the value of σ_u^2. Therefore, a close estimate of σ_u^2 to the true value is more important than having a good estimate of underlying covariance structure. Hence, we recommend that a sensitivity analysis be used in practice.

To evaluate the performance of the proposed method under small samples, we also conducted a simulation with a sample size of 50. In this case, the estimates obtained from Method I are slightly upwardly biassed with higher standard errors. However, Method II adjusts for bias quite well and provides reliable estimates of standard errors. In general, considering all the simulation scenarios, the transformation method (Method II) outperforms the method of moments (Method I).

We also run another set of simulations to ascertain whether the spatial configuration of the point locations is an important feature in determining the effectiveness of bias correction. We generated data based on a cubic function with Gaussian noise in an 80×80 grid rather

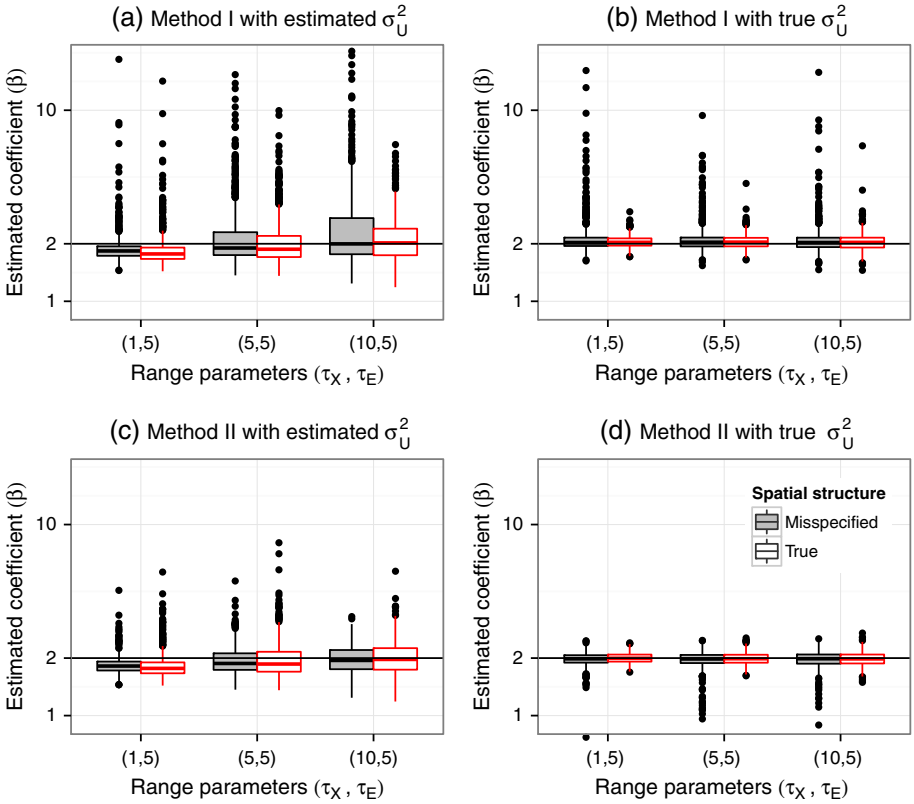

Figure 2. Distribution of estimated coefficient when estimated and true values of σ_u^2 used with Methods I (a–b) and II (c–d) under different range parameter combinations with true and misspecified covariate structures

than uniformly distributed within the grid. Our simulation results show that the spatial configuration affects the estimates of measurement error variance. However, our methods are quite robust in case of misspecification of underlying spatial configuration; hence, adjust bias very well if true measurement error variances are known (results not shown).

6. ANALYSIS OF ISCHAEMIC HEART DISEASE DATA

Data on IHD were collected from all hospitals in New South Wales, Australia, between 1 July 1994 and 30 June 2002. A detailed description of the data has been given elsewhere (Burden *et al.*, 2005). Briefly, patients who were admitted to the hospitals via the emergency room and discharged with a diagnosis of IHD were considered as acute IHD cases. Data also include patient age, gender and geographic location reported via postcodes of residence. Data from 579 postcodes were included in the analysis. IHD event data were linked with the census data, which contain age-specific and gender-specific population counts. SEIFA scores and centroid coordinates (latitude and longitude) for each postcode were obtained from the Australian Bureau of Statistics. Because temporal patterns were not our main concern in this study, we averaged the 8-year SEIFA scores and aggregated values of the population size and number of IHD admitted cases for each postcode. We then calculated age–sex-adjusted standardized incidence ratios by dividing the observed number of IHD cases by the age–sex-adjusted expected IHD cases (Breslow and Day, 1987).

Li *et al.* (2009) analysed square root-transformed standard mortality ratios to make them more normally distributed. However, we found that untransformed standardized incidence ratio values more closely approximated the normal distribution, and hence, we did not transform them. We fit model (1) assuming an exponential correlation structure for data observed for each postcode, with distance based on the latitude and longitude of each postcode centroid. As Burden *et al.* (2005) noted, the principal component analysis that was used to derive the SEIFA score only accounts for about 30% of the total variation of the component used. Therefore, it is likely that the SEIFA score is subject to substantial measurement error. We standardized the SEIFA scores to have a mean of zero.

The results of our analysis are given in Table 2. The naïve analysis ignoring spatial correlation suggests a significant protective effect associated with higher SEIFA values ($\hat{\boldsymbol{\beta}}_{SEIFA} = -0.062$ with $SE = 0.014$). Analysis via a linear mixed model accounting for spatial correlation also suggests that the effect is very strong ($\hat{\boldsymbol{\beta}}_{SEIFA} = -0.141$ with $SE = 0.015$). However, the magnitude of the effect is much larger.

Table 2. Analysis of ischaemic heart disease data in New South Wales, Australia, under different specifications of measurement error

	Estimates for SEIFA	
Methods	$\hat{\boldsymbol{\beta}}$	$SE(\hat{\boldsymbol{\beta}})$
Ignoring measurement error		
OLS	−0.062	0.014
LME with spatial correlation	−0.141	0.015
Accounting for measurement error bias		
Method I	−0.377	0.041
Method II	−0.278	0.015

SEIFA, Socioeconomic Indicators for Areas.

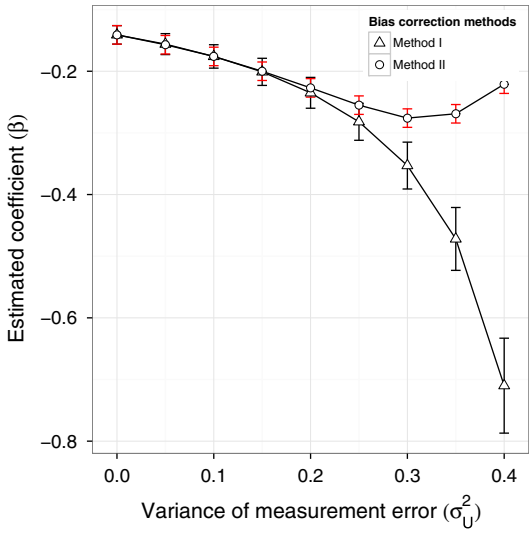

Figure 3. Sensitivity analysis for ischaemic heart disease data. The assumed measurement error variance varied between 0 (naïve) and 0.40

We applied our bias correction methods on the result obtained from the linear mixed model. The linear mixed model of SEIFA based on an intercept-only model with assumed exponential spatial correlation suggests the estimate of measurement error variance as 0.28. Both methods suggest a strongly significant effect of SEIFA ($\hat{\boldsymbol{\beta}}^{adj}_{SEIFA} = -0.377$ and -0.278 with $SE = 0.041$ and 0.015, respectively). A large difference in the estimated standard error for Methods I and II is observed. The estimated standard error for Method I is given by $\hat{Var}(\hat{\beta}^{adj}_1) = \hat{\rho}^{-2}\sigma^2_*$ which may be subject to produce bias estimates for a given value of attenuation factors, $\hat{\rho}$. As the estimated standard error for Method I does not account for the variability of the attenuation factor. In practice, a bootstrap procedure can be used to calculate the standard error for Method I. We implement a block bootstrap procedure by leaving one block at each iteration. Blocks are automatically selected using the cluster separation method *clara* (Kaufman and Rousseeuw, 2005) in *R* (R Core Team, 2013). Specifically, this method selects k representative objects in the data set, where k is the number of clusters. The remaining objects are then assigned to the nearest representative object to form a cluster. The representative objects are selected in such a way that the average distance of the representative objects to all other objects in the same cluster is minimized. Our results show that the difference in estimated standard error reduces with large number of blocks, while the estimated standard error for Method II remains unchanged (results not shown).

Because we do not have a validation data set and thus cannot test the assumption underlying the bias correction methods, we conduct a sensitivity analysis to help in the interpretation of our results. We conducted sensitivity analysis by varying measurement error variance, σ^2_U from 0.0 (naïve) to 0.40. The result of the sensitivity analysis is presented in Figure 3.

As measurement error variance σ^2_U increases, the estimates obtained by method of moments also decrease. The estimates obtained using transformation methods also decrease until the assumed measurement error variance is less than the estimated measurement error variance and then increase. We note that the transformation method appears to give stable results over the range of σ^2_U.

7. DISCUSSION

In this paper, we have developed a framework to quantify the bias induced in estimated regression coefficients when covariates are measured with error in spatial regression settings. Both analytic and simulation results suggest that naïve analysis that ignores measurement error will attenuate estimated regression coefficients towards the null hypothesis of no effect. Our results extend classical measurement error theory in

that the amount of attenuation depends on the degree of spatial correlation in both the covariate of interest and the assumed random error from the regression model. These results explain why the results from spatial regression modelling are often so sensitive to the assumed model error structure. We proposed two different strategies to obtain consistent estimates of the regression coefficients of interest in the presence of covariate measurement error. These strategies include the following: (i) *post hoc* adjustment of estimated regression coefficient via an estimate of the attenuation factor; and (ii) a linear transformation of the error-prone covariates that can then be analysed to yield consistent results. We found that both methods perform well, although the second method tends to be less variable and hence preferable in practice. We present formulae for the standard errors of the adjusted estimated regression coefficients, although these do not fully account for the uncertainty associated with the estimation of unknown parameters. In practice, a bootstrap procedure can be used to obtain appropriate standard errors. We illustrated the proposed approaches using the analysis of IHD data. There are a number of areas where future study would be useful.

Our analytic results are similar to those of Li *et al.* (2009) who also consider asymptotic bias associated with spatial regression analysis involving covariate measurement error. They also propose an adjusted analysis based on an EM algorithm. However, their approach is difficult to apply, especially for large data sets. In contrast, our proposed approaches can be easily implemented using readily available packages such as *lme* in R. While Li *et al.* (2009) demonstrate via simulation that their method works well, they use the true values of the measurement error variances and did not consider the setting where measurement error variances are estimated. While our simulations confirm the reliability of our proposed methods in settings where the measurement error variance can be assumed known, we also suggest an approach to estimating the measurement error variance. As in the classical measurement error context, estimation of measurement error variance requires either additional assumptions or additional information such as validation or replicate data. We showed how the measurement error variance could be estimated under the assumption that the true covariate of interest is smooth and hence that any estimated nugget effect can be interpreted as a measurement error. As expected, our simulations suggest that the performance of our proposed bias correction methods declines when the measurement error variance is estimated instead. Method I performs more poorly than Method II because the latter requires the estimation of fewer model parameters. Our results suggest that having some knowledge of the magnitude of measurement error is important, and in practice, we suggest the use of a sensitivity analysis that varies the assumed values of the measurement error variance over a feasible range.

One observation from our simulation is that the use of an estimated measurement error variance from the data leads to underestimation of the regression coefficients when there is minimal spatial correlation in the covariate. This makes sense because most of the covariate variability will be absorbed into the estimated nugget effect. As expected, we found that the situation improved when we used a smaller grid size (see also Bell and Grunwald 2004). Use of an estimated measurement error variance also led to much greater discrepancies between the average of our estimated standard errors and the empirical standard errors derived from the simulation. In contrast, these were much closer when the true values of the measurement error variances were used. These observations suggest that having knowledge of the true measurement error variance is crucial not only in obtaining consistent estimates of the regression coefficients but also in terms of estimating standard errors and conducting appropriate inference. Again, we recommend the use of a sensitivity analysis in practice.

Our heart disease example demonstrated a substantial increase in the rates of IHD as the level of SEIFA measured at the postcode level decreased. The magnitude of the effect increased after adjusting for measurement error. Our results are consistent with the social epidemiology literature (see systematic review by Pickett and Pearl 2001) that suggests that low socio-economic status leads to increased rates of a wide variety of health outcomes. While it is tempting to interpret these results at an individual level, it is important to remember that doing so may result in ecological bias (Sheppard, 2003). Prentice and Sheppard (1995) showed that using group-level covariates in the analysis reduces the effects of error in the measurement of covariates at the individual level. However, Greenland (2001) and Jackson *et al.* (2006) noted that ecological covariates are subject to non-random survey errors and may not be addressed by aggregation of group-level analysis of covariates. Moreover, in many research areas, group-level data are the only available source for analysis. Air pollution epidemiology provides a classic example, because individual measurements of air pollution studies are rarely collected and instead are estimated based on neighbourhood monitoring and other sources (Sheppard *et al.*, 2012). Consequently, air pollution exposures are typically measured with error, and it would be useful to consider the impact of this error on subsequent effect size estimates.

In our simulation, we have considered only a single covariate measured with error in a spatial linear mixed model with Gaussian error. It would be of interest to explore the effect of covariate measurement error in the presence of multiple covariates and also omitted covariates. Future work can also be performed on extending our formulation to the spatial generalized linear mixed model with non-Gaussian outcomes. However, such explorations are beyond the scope of this present paper.

In light of the increasing popularity of multi-level models that include both individual and area-specific covariates, it is important that practitioners are aware of the importance not only of careful modelling of the mean function but also of accounting for measurement error and appropriate spatial structure of their data.

ACKNOWLEDGMENTS

The authors thank an unknown reviewer for helpful comments on the initial draft of this paper. HDB was partially supported as a visitor at the School of Mathematical Sciences, University of Technology, Sydney, and by grants NSF DMS-1308400 and NIH P01-CA142538. LR and HH were supported by the University of Technology, Sydney and by the ARC Centre of Excellence for Mathematical & Statistical Frontiers (ACEMS). The authors thank the NSW Ministry of Health for making the data available.

REFERENCES

Bell ML, Grunwald GK. 2004. Mixed models for the analysis of replicated spatial point patterns. *Biostatistics* **5**(4):633–648.

Breslow NE, Day NE. 1987. *Statistical methods in cancer research. Volume II–The design and analysis of cohort studies*. International Agency for Research on Cancer. Oxford University Press: New York, U.S.A.

Breslow NE, Clayton DG. 1993. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**(421):9–25.

Burden S, Guha S, Morgan G, Ryan L, Sparks R, Young L. 2005. Spatio-temporal analysis of acute admissions for ischemic heart disease in NSW, Australia. *Environmental and Ecological Statistics* **12**(4):427–448.

Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. 2006. *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman and Hall/CRC: Boca Raton, Florida, U.S.A.

Carroll RJ, Chen R, George EI, Li TH, Newton HJ, Schmiediche H, Wang N. 1997. Ozone exposure and population density in Harris County, Texas. *Journal of the American Statistical Association* **92**(438):392–404.

Clayton DG, Bernardinelli L, Montomoli C. 1993. Spatial correlation in ecological analysis. *International Journal of Epidemiology* **22**(6):1193–1202.

Cook DG, Pocock SJ. 1983. Multiple regression in geographical mortality studies, with allowance for spatially correlated errors. *Biometrics*:361–371.

Elliot P, Wakefield JC, Best NG, Briggs DJ. 2000. *Spatial Epidemiology: Methods and Applications*. Oxford University Press: New York, U.S.A.

Elliott P, Wartenberg D. 2004. Spatial epidemiology: current approaches and future challenges. *Environmental Health Perspectives* **112**(9):998–1006.

Fuller WA. 1987. *Measurement Error Models*, Wiley Series in Probability and Statistics. John Wiley & Sons: New York, U.S.A.

Greenland S. 2001. Ecologic versus individual-level sources of bias in ecologic estimates of contextual health effects. *International Journal of Epidemiology* **30**(6):1343–1350.

Gryparis A, Paciorek CJ, Zeka A, Schwartz J, Coull BA. 2009. Measurement error caused by spatial misalignment in environmental epidemiology. *Biostatistics* **10**(2):258–274.

Guha S, Ryan L, Morara M. 2009. Gauss–Seidel estimation of generalized linear mixed models with application to Poisson modeling of spatially varying disease rates. *Journal of Computational and Graphical Statistics* **18**(4):818–837.

Jackson C, Best N, Richardson S. 2006. Improving ecological inference using individual-level data. *Statistics in Medicine* **25**(12):2136–2159.

Kaufman L, Rousseeuw PJ. 2005. *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley Series in Probability and Statistics. John Wiley & Sons: Hoboken, New Jersey, U.S.A.

Li Y, Tang H, Lin X. 2009. Spatial linear mixed models with covariate measurement errors. *Statistica Sinica* **19**(3):1077–1093.

Molitor J, Jerrett M, Chang C-C, Molitor N-T, Gauderman J, Berhane K, McConnell R, Lurmann F, Wu J, Winer A, Thomas D. 2007. Assessing uncertainty in spatial exposure models for air pollution health effects assessment. *Environmental Health Perspectives* **115**(8):1147–1153.

Pickett KE, Pearl M. 2001. Multilevel analyses of neighbourhood socioeconomic context and health outcomes: a critical review. *Journal of Epidemiology and Community Health* **55**(2):111–122.

Pinheiro J, Bates D, DebRoy S, Sarkar D, R Core Team: nlme: linear and nonlinear mixed effects models. R package version 3.1-109

Prentice RL, Sheppard L. 1990. Dietary fat and cancer: consistency of the epidemiologic data, and disease prevention that may follow from a practical reduction in fat consumption. *Cancer Causes & Control* **1**(1):81–97.

Prentice RL, Sheppard L. 1995. Aggregate data studies of disease risk factors. *Biometrika* **82**(1):113–125.

R Core Team: R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria

Ruppert D, Wand MP, Carroll RJ. 2009. Semiparametric regression during 2003–2007. *Electronic Journal of Statistics* **3**:1193–1256.

Schwartz J, Coull BA. 2003. Control for confounding in the presence of measurement error in hierarchical models. *Biostatistics* **4**(4):539–553.

Sheppard L. 2003. Insights on bias and information in group-level studies. *Biostatistics* **4**(2):265–278.

Sheppard L, Burnett RT, Szpiro AA, Kim S-Y, Jerrett M, Pope III, CA, Brunekreef B. 2012. Confounding and exposure measurement error in air pollution epidemiology. *Air Quality, Atmosphere & Health* **5**(2):203–216.

Szpiro AA, Sheppard L, Lumley T. 2011. Efficient measurement error correction with spatially misaligned data. *Biostatistics* **12**(4):610–623.

Waller LA, Gotway CA. 2004. *Applied Spatial Statistics for Public Health Data*, Vol. 368. John Wiley & Sons: Hoboken, New Jersey, U.S.A.

Wansbeek TJ, Meijer E. 2000. *Measurement Error and Latent Variables in Econometrics*, Vol. 37. Elsevier Amsterdam: North-Holland.

Wood S. 2006. *Generalized Additive Models: An Introduction With R*. Chapman and Hall/CRC: Boca Raton, Florida, U.S.A.

Xia H, Carlin BP. 1998. Spatio-temporal models with errors in covariates: mapping Ohio lung cancer mortality. *Statistics in Medicine* **17**(18):2025–2043.

Zeger SL, Thomas D, Dominici F, Samet JM, Schwartz J, Dockery D, Cohen A. 2000. Exposure measurement error in time-series studies of air pollution: concepts and consequences. *Environmental Health Perspectives* **108**(5):419–426.

Zheng Y, Zhu J. 2012. On the asymptotics of maximum likelihood estimation for spatial linear models on a lattice. *Sankhya A* **74**(1):29–56.

APPENDIX

The ordinary least squares estimate of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}^{ols} = \left(\boldsymbol{W}_*^T \boldsymbol{W}_*\right)^{-1} \boldsymbol{W}_*^T \boldsymbol{Y} \tag{A1}$$

with $\boldsymbol{W}_*$ defined in the text. Under the true model, $\boldsymbol{Y} = \boldsymbol{X}_*\boldsymbol{\beta} + \boldsymbol{\epsilon}$, we have

$$\begin{aligned}
\hat{\boldsymbol{\beta}}^{ols} &= \left(\boldsymbol{W}_*^T \boldsymbol{W}_*\right)^{-1} \boldsymbol{W}_*^T \boldsymbol{Y} \\
&= \left(\boldsymbol{W}_*^T \boldsymbol{W}_*\right)^{-1} \boldsymbol{W}_*^T \left(\boldsymbol{X}_*\boldsymbol{\beta} + \boldsymbol{\epsilon}\right) \\
&= \left(\boldsymbol{W}_*^T \boldsymbol{W}_*\right)^{-1} \boldsymbol{W}_*^T \boldsymbol{X}_*\boldsymbol{\beta} + \left(\boldsymbol{W}_*^T \boldsymbol{W}_*\right)^{-1} \boldsymbol{W}_*^T \boldsymbol{\epsilon} \\
&= \left(\frac{\boldsymbol{W}_*^T \boldsymbol{W}_*}{n}\right)^{-1} \left(\frac{\boldsymbol{W}_*^T \boldsymbol{X}}{n}\right) \boldsymbol{\beta} + \left(\frac{\boldsymbol{W}_*^T \boldsymbol{W}_*}{n}\right)^{-1} \left(\frac{\boldsymbol{W}_*^T \boldsymbol{\epsilon}}{n}\right)
\end{aligned}$$

Now under certain regularity conditions (Zheng and Zhu, 2012) and by the weak law of large numbers, $\left(\boldsymbol{W}_*^T \boldsymbol{W}_*/n\right) \xrightarrow{p} cov\left(\boldsymbol{W}_*\right)$, $\left(\boldsymbol{W}_*^T \boldsymbol{X}_*/n\right) \xrightarrow{p} cov\left(\boldsymbol{X}_*\right)$ and $\left(\boldsymbol{W}_*\boldsymbol{\epsilon}/n\right) = \left(\boldsymbol{X}_*\boldsymbol{\epsilon}/n + \boldsymbol{U}_*\boldsymbol{\epsilon}/n\right) \xrightarrow{p} 0$. It follows that $\tilde{\boldsymbol{\beta}}^{ols} \xrightarrow{p} \left[cov\left(\boldsymbol{W}_*\right)\right]^{-1} cov\left(\boldsymbol{X}_*\right)\boldsymbol{\beta}$. Because $\boldsymbol{W}_*$ and

$\boldsymbol{X}_*$ have first column equal to $\mathbf{1}$ corresponding to the intercept of the model and assuming $\mu_X = 0$, it follows that

$$\tilde{\boldsymbol{\beta}}^{ols} \xrightarrow{p} \begin{pmatrix} 1 & 0 \\ 0 & trace\,(\boldsymbol{\Sigma}_X + \boldsymbol{\Sigma}_U) \end{pmatrix}^{-1} \begin{pmatrix} 1 & 0 \\ 0 & trace\,(\boldsymbol{\Sigma}_X) \end{pmatrix} \boldsymbol{\beta} = \begin{pmatrix} 1 & 0 \\ 0 & \rho^{ols} \end{pmatrix} \boldsymbol{\beta},$$

where $\rho^{ols} = trace\,(\boldsymbol{\Sigma}_X)\,/trace\,(\boldsymbol{\Sigma}_X + \boldsymbol{\Sigma}_U)$.

